TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 40, 131–150 (1991)

An Evaluation of Delphi

FRED WOUDENBERG

ABSTRACT

The literature concerning quantitative applications of the Delphi method is reviewed. No evidence was found to support the view that Delphi is more accurate than other judgment methods or that consensus in a Delphi is achieved by dissemination of information to all participants. Existing data suggest that consensus is achieved mainly by group pressure to conformity, mediated by the statistical group response that is fed back to all participants.

Introduction

Human judgment is necessary in situations of uncertainty and, because these situations abound, is very much relied upon. However, numerous reports about flaws in human judgment and its inferiority to more formal methods of judgment have appeared [1–6]. Research since the beginning of this century has sought to discover the factors responsible for the shortcomings in human judgment, with the aim of developing more accurate judgment methods. In reviewing these studies, several authors [7–12] draw the same three conclusions:

1. A statistical aggregate of several individual judgments is more accurate than judgment of a random individual. Some authors [10, 12] refine this by noting that this holds only for tasks of simple to intermediate difficulty.
2. Judgments resulting from interacting groups are more accurate than a statistically aggregated judgment. Also, interaction leads to stronger agreement.
3. Unstructured, direct interaction still has disadvantages that lead to suboptimal accuracy of judgments.

Starting from these three premises, the most logical next step is to develop judgment methods that possess all the advantages, but not the disadvantages, of unstructured, direct interaction. Many of these methods have been developed. In most, one tries to overcome the disadvantages of unstructured, direct interaction by structuring the interaction between participants. The nominal group technique, developed by van de Ven and Delbecq [13], is the most widely known structured, direct interaction method. In other procedures, interaction is also structured, but no allowance is made for direct interaction. The best-

FRED WOUDENBERG received a PhD in experimental psychology in 1989.

Address reprint requests to Dr. Fred Woudenberg, Scientific Department, Environmental Health Service Rotterdam, PO Box 70032, 3000 LP Rotterdam, The Netherlands.

 0040-1625/91/$3.50

TABLE 1
Judgment Methods in Order of Increasing Accuracy Based on Literature Review or Expectation

Traditional methods			Newly developed methods	
Random individual	Staticized group	Unstructured, direct interaction	Structured, direct interaction	Delphi
Accuracy				

Solid line, literature review; dashed line, expectation.

known structured, indirect interaction method is the Delphi technique, which is the focus of the present article.

Based on the above-mentioned literature reviews and the expectations derived from them [see, e.g., 14], the different judgment methods can be put on a scale of increasing accuracy (Table 1). In this article the expected high relative accuracy of Delphi is evaluated. Both the accuracy of the Delphi method as a whole and the contribution of the separate Delphi characteristics to its accuracy will be discussed. In addition, the reliability of Delphi and, briefly, its capacity to induce consensus are evaluated.

History of Delphi

The first experiment using a Delphi methodology was performed in 1948 to improve betting scores at horse races [15]. The name "Delphi" was coined by Kaplan [see 16], a philosopher working for the Rand Corporation who headed a research effort directed at improving the use of expert predictions in policy-making. Kaplan et al. [17] found that unstructured, direct interaction did not lead to more accurate predictions than statistical aggregation of individual predictions. To make better use of the potential of group interaction, Gordon, Helmer, and Dalkey, also at the Rand Corporation, developed the Delphi method in the 1950s. Between 1950 and 1963 they conducted 14 Delphi experiments, but, as a consequence of its military character, all this work was secret. The first article describing some of this research was published in 1963 [18]. In 1964, Gordon and Helmer published an article that roused worldwide interest in Delphi [19].

Delphi was developed as a method to increase the accuracy of forecasts. Many other types of Delphi have been derived from the original method:

- Delphi to estimate unknown parameters [20]
- Policy Delphi [21]
- Decision Delphi [22]

A Delphi has even been used to compile an epidemiological dictionary [23].

In the 1950s and 1960s, Delphi was used mainly to make quantitative assessments (forecasting dates and estimating unknown parameters). In the 1970s, stress was more and more put on the educational and communicational possibilities of Delphi [24–28], although Dalkey [29] had mentioned these possibilities in 1967. Some authors [30, 31] began to call Delphi a "communication device" and measured its success qualitatively as the satisfaction of the participants with the method instead of quantitatively as accuracy [32]. The evaluation of Delphi in the present article is restricted to the quantitative Delphi, and the conclusions pertain only to this form.

Characteristics of Delphi

The characteristics of Delphi as it was originally developed are

Anonymity. Participants, mostly experts, are approached by mail or computer.

Iteration. There are several rounds. The first round can be inventory, in which participants are asked for events to be forecasted or parameters to be estimated. In subsequent rounds, participants are asked to give quantitative estimates about dates of future events or values of unknown parameters. The number of rounds is fixed in advance or determined according to a criterion of consensus in the group of participants or stability in individual judgments.

Feedback. The results of an eventual first inventory round are clustered and sent back to all participants. In the first estimation round, participants give their quantitative estimates. Before the second and subsequent estimation rounds, the results of the whole group on the previous round are fed back in a statistical format (measure of central tendency plus variance) to all participants. On the second and subsequent estimation rounds, participants making judgments that deviate from the first-round group score according to a fixed criterion are asked to give arguments for their deviating estimates. Before the third and subsequent estimation rounds, these arguments are, along with the statistical results, fed back to all participants.

Many variations on this standard method have been used. Delphis with partial anonymity have been conducted [22]. The number of rounds varies from two [33] to ten [34]. Delphis without a first inventory round are often used to save time. If an inventory is necessary, it is done by other means, such as interviewing key persons. Statistical feedback in a Delphi can vary from a single number [35] to complete distributions [36]. Feedback of arguments is rarely given.

Evaluation of Delphi

The accuracy[1] and reliability[2] of a judgment method are difficult to evaluate. The reason for this is that judgments cannot be equated to measurements. A measurement can be partitioned into a true score and an error component. The error component can be regarded as consisting of a number of random variables [37]. These random variables tend to cancel each other out in the long run, giving the error component an expectation of zero.

A judgment can also be thought of as consisting of a true score and an error component. However, the error component cannot be regarded as consisting of random variables. More realistically, the error component in a judgment can be thought of as being influenced by person- and situation-specific factors. This means that with a judgment method there is ample opportunity for bias, and this bias can vary from application to application of the method. As a consequence, every new application of the method can

[1]Accuracy is meant here as the correspondence between the judgment and the true value; in statistical terms, it is designated as external validity.

[2]Reliability is the certainty with which an instrument (for instance, a judgment method) reflects true scores and not random errors. Reliability indicates the reproducibility of an instrument. High reliability implies that measurements (judgments) reflect the true score, but it does not guarantee the true score is correct and does not contain systematic error. This is reflected in the common remark that high reliability is a necessary—but not sufficient—condition for high accuracy.

TABLE 2
Comparison of the Accuracy of Judgment Methods with Delphi as Reference

Study	Staticized group	Unstructured, direct interaction	Structured, direct interaction	Delphi
Campbell [41]		2		1
Pfeiffer [43]	2			1
Dalkey [20]		2		1
Farquhar [51]	1			2
Gustafson et al. [52]	2	2	1	3
Sack [44]		2		1
Brockhoff [59]		1		1
Ford [49]	1			2
Seaver [55]	1	2	2	2
Miner [53]			1	1/2
Moskowitz and Bajgier [54]		1		2
Rohrbaugh [12]			1	1
Fischer [57]	1	1	1	1
Boje and Murnighan [48]	1		3	2
Riggs [33]		2		1
Parenté et al. [42]	2			1
Erffmeyer and Lane [14]		2	3	1

Methods are ranked from most accurate (1) to less accurate (2, 3).

be seen as a new measuring instrument. Evaluation of the accuracy and reliability of Delphi, being a judgment method, is therefore seriously hampered by the possible influence of person- and situation-specific biases.[3]

Accuracy of the Delphi Method as a Whole

Because it is difficult to evaluate the accuracy of a judgment method, it is not surprising that the accuracy of the Delphi has been falsely inferred from other criteria, such as consensus [38], the log-normality of first estimates [39], and the relation between remoteness and precision of a forecast [40]. The most feasible way to evaluate the accuracy of Delphi is to compare it directly to other judgment methods in the same situation. The main sources of bias that remain are order effects when the same participants are used with the different methods and person-specific effects when different groups of participants are used. These effects can be minimized with proper randomization.

A number of studies have been reported in which accuracy was evaluated by comparing Delphi directly to other methods (see Table 2). In most studies no statistical comparison between methods was made. Therefore, the accuracy of the investigated methods was rank ordered from most to least accurate. Table 3 lists all 26 possible pairwise comparisons between methods of the 17 mentioned studies. A slight—but not unequivocal—indication for Delphi's expected higher accuracy as compared to unstructured, direct interaction can be observed. A similarly unequivocal suggestion can be found in Delphi's lower accuracy as compared to the staticized group. The two suggestions taken together (Delphi being more accurate than unstructured, direct interaction, but less accurate than a staticized group) are not easy to interpret and can even be called anomalous.

[3]Note that not only is the evaluation of the accuracy and reliability of the Delphi method hampered by person- and situation-specific biases, but also its validation and standardization.

TABLE 3
Pairwise Comparisons of the Accuracy of Judgment Methods Based on the Results of Table 2

		Delphi		
		More accurate	Equally accurate	Less accurate
Staticized group	vs Delphi	2	1	5
Unstructured, direct interaction	vs Delphi	5	3	2
Structured, direct interaction	vs Delphi	2	4	2
All methods	vs Delphi	9	8	9

In general (see the *Introduction*), it has been found that unstructured, direct interaction gives more accurate results than does a staticized group. The present suggestions imply that a staticized group is more accurate than unstructured, direct interaction. Unfortunately, very few direct comparisons between unstructured, direct interaction and a staticized group were made in the reviewed studies. In three comparisons, a staticized group was once more accurate and twice equally accurate, slightly supporting the anomalous conclusion. The comparison between Delphi and structured, direct interaction suggests that there is no difference in accuracy. Also, the comparisons between Delphi and all other methods (meaningful because Delphi has been proposed as the most accurate judgment method available) show no difference.

Taken together, the reviewed studies do not offer easily interpretable conclusions or unequivocal outcomes of comparisons. A closer scrutiny of the separate studies does not offer a clue as to the factors determining these unequivocal results. A criticism pertaining to all the studies is that experiments were conducted in a laboratory, while the quiescence of the private environment has been mentioned as a distinct advantage of Delphi. Also, in almost no study were expert participants used. Furthermore, in most studies no arguments of deviating judges were asked for and fed back to the entire group.

More specific criticisms can be leveled against the studies, including those seven [14, 20, 33, 41–44] that found Delphi to be the most accurate method. The experiments Campbell [41] describes in his thesis are often cited as evidence of Delphi's higher accuracy as compared to unstructured, direct interaction. Sackman [45], although calling Campbell's experiments well conducted, criticizes them because the unstructured, direct interacting control group was not really unstructured, but "force-fitted into a Delphi-type format." Although Campbell did this to make comparison with Delphi feasible, Sackman concludes that it obstructs comparison because the participants in Campbell's interacting group did not get enough opportunity for spontaneous interaction, and therefore this group has to be regarded as seriously disadvantaged in comparison to Delphi.

Pfeiffer [43] reported Delphi to be more accurate than what seems to be a staticized group response for 13 of 16 short-term economic indicators. This result was not obtained by Pfeiffer himself. In his book, Pfeiffer describes these results in two short paragraphs. He only writes that the experiments were conducted at the University of California in Los Angeles. He seems to refer to the experiments performed by Campbell. But Campbell's research concerned the comparison between Delphi and unstructured, direct interaction. It is not clear whether Pfeiffer is wrongly referring to this research or rightly referring to a separate part of this research that has not been described elsewhere.

The experiments by Dalkey [20, 46, 47] probably most strongly contributed to the view that the accuracy of judgments can be increased by using the Delphi method. In a series of 11 experiments, Dalkey asked students to answer almanac-type questions, such

as "In what year was nylon invented?" and "What was the circulation of *Playboy* magazine as of January 1, 1966?" Dalkey's experiments are strongly criticized for several reasons. First, it should be noted that the judgments in Dalkey's experiments did not always become more accurate over rounds. Improvement occurred in two-thirds of the questions, and deterioration was found in the remaining one-third. The tasks Dalkey used were very simple, and some authors [48] considered them unrepresentative of most judgmental problems encountered in real life. Also, the nonparametric statistical tests Dalkey used to evaluate his results are criticized for their simplicity [49]. Dalkey only recorded whether improvement over rounds did or did not occur, independent of the absolute change in accuracy. As a result, very small changes in accuracy can attain much weight. A striking example of this is found in a study performed by Brown and Helmer [50], using Dalkey's method of analysis. Here, increased accuracy over rounds was also found for two-thirds (13 of 20) and decreased accuracy for one-third (7 of 20) of the questions. Of the 13 questions improving in accuracy over rounds, the improvement was 0–0.1 standard scores for five questions. 0.1–0.2 standard scores for four questions, 0.2–0.3 standard scores for three questions, and almost 0.5 standard score for only one question. That these increases are very modest can be seen in the comparison of last-round Delphi scores with the first-round scores of a random individual judge, which can be regarded as an absolute minimal baseline. Delphi scores were more accurate on a median of 12.5 of 20 questions, while the random judge was more accurate on the remaining 7.5 questions. An even less favorable picture appears when interquartile distances are considered. In the first round, 13 of the 20 interquartile distances covered the true value. Because of convergence, only seven values fell in the interquartile distance on the last round.

An unpublished report by Sack [44] is cited by Riggs [33] in two sentences. The only information given is that Delphi was more accurate in short-term forecasting than an unstructured, direct interacting group, but not at the stated level of statistical significance. Riggs's own study [33], in which Delphi was found to be more accurate than unstructured, direct interaction, made use of a modified Delphi. Participants were not anonymous and only two rounds were run.

In the study by Parenté et al. [42], Delphi was more accurate than a staticized group, but the comparison between Delphi and the staticized group was done with different groups of participants in two different situations. First, for ten events, a group of 300 students forecasted if and (if so) when the events would occur during the next months. Three groups of 100 students each made forecasts weekly during a time span of one, two, or three months. A staticized group result was calculated based on two events that occurred in the designated time period. Following this, a Delphi experiment was conducted. This time, 80 new students made forecasts about 30 new events. Again, students had to make a forecast every week, all for a time span of two months. Of the 30 events, six occurred in these two months and these six events were used in the analysis. It is clear that the different groups of participants and events, as well as the different number of participants and events, seriously hinders the comparison between the staticized group and the Delphi.

The study by Erffmeyer and Lane [14] compared four procedures using the NASA "lost on the moon" exercise: (a) unstructured, direct interaction, (b) consensus group, (c) structured, direct interaction, and (d) Delphi. Participants in the consensus group engaged in unstructured direct interaction, but followed guidelines to resolve conflicts. Participants had to rank 15 items of equipment in terms of importance for the survival of a shipwrecked crew on the moon. The results of four groups, each using one of the four different procedures, were compared to the "correct" rank order assigned by NASA

experts. The Delphi group was the most accurate, followed by the consensus group and the unstructured, direct interacting group; the structured, direct interacting group was the least accurate. These results can, of course, be questioned because the correct rank order is unknown. There is no way to tell whether the NASA experts' rank order is correct. It would be interesting to know if the experts themselves would rank order the items identically under the four different procedures.

The above-mentioned criticisms of studies reporting Delphi to be more accurate than one or more other judgment procedures can be and have been used against Delphi. The seven studies [48, 49, 51–55] in which Delphi was found to be less accurate than at least one other method can, however, likewise be criticized.

In the study by Farquhar [51], Delphi was found to be less accurate than unstructured, direct interaction in two separate experiments. But the interpretation of these results is hampered by the small and inequal number of participants. The Delphi groups had nine and four participants, and the unstructured, direct interacting groups had five and three. Group size can have a profound effect on the accuracy of judgments [20, 56].

Gustafson et al. [52] compared Delphi with a staticized group, with unstructured, direct interaction, and with structured, direct interaction. Delphi was the least accurate method. These results have been questioned on two grounds. Fischer [57] contends that the dependent variable used by Gustafson et al. exaggerates small differences, and he reanalyzes the results with another dependent variable. The outcome of this analysis is that no differences between methods are found. Van de Ven and Delbecq [58] ascribe Gustafson et al.'s negative findings with Delphi to the specific variant of Delphi used. This was not a real Delphi, but a derived method called "estimate–feedback–estimate." Small groups of four participants sat together and received written feedback. The unnaturalness of written feedback in a gathered group with the prohibition of direct contact could, according to van de Ven and Delbecq, have induced negative social affiliation with detrimental effect on the results.

In the study by Ford [49], a staticized group was more accurate than two versions of Delphi. But the absolute differences in accuracy were small. The only striking result was that over four rounds Delphi estimates became slightly less accurate and staticized group estimates slightly more accurate. Ford's study can be criticized for his use of a within-subjects design in which four groups of subjects all participated in four different judgment methods. Although Ford neatly randomized the order in which the four experimental groups participated in the four methods by making use of a Latin square, his results do suggest order effects. Randomizing order allows for testing order effects, but it is no guarantee that order effects will not occur. For instance, having a Delphi (in which feedback is provided) before a staticized group method (iteration of individual judgments in which feedback is not provided) cannot be expected to be counterbalanced by giving another group of subjects the staticized group method before the Delphi.

The study by Seaver [55] compared five methods (no interaction; unstructured, direct interaction; two methods using structured, direct interaction; and Delphi) under six aggregation procedures. Averaged over all aggregation procedures, all kinds of interaction tended to decrease the accuracy of probabilistic answers to two-alternative general knowledge questions. The design and methodology of the study were quite complex. Ten groups of four persons each answered questions before and after interaction (except in the no-interaction condition). Five sets of questions and the five procedures were balanced using a Greco-Latin square design. Just as in the study by Ford [49], the randomization of order does not imply that order effects do not occur. It cannot be determined whether Seaver's data, like those of Ford, suggest order effects, because Seaver's paper does not

give the worked-out order of procedures. In sharp contrast to the statistical sophistication of the study, the interaction procedures Seaver used were of rudimentary form. Delphi was barely Delphi-like. Only two rounds were used. Before the second round the experimenter personally informed each group member about the other members' assessed probabilities and the reasons for them. Therefore, individual scores, but not a group score, were fed back to all participants.

The study by Miner [53] reports Delphi to be significantly less accurate than a structured, direct interaction method that was the focus of the article. But there was no difference between Delphi and the most widely known structured, direct interaction method, the nominal group technique. Erffmeyer and Lane [14] criticize the type of Delphi used by Miner, primarily for the lack of physical separation of participants.

Moskowitz and Bajgier [54] found Delphi to be less accurate than unstructured, direct interaction. The comparison of both methods was, however, not the primary object of the study. Little attention was paid to the characteristics of the Delphi. Participants were not anonymous and feedback of individual scores instead of group scores was given.

In the study by Boje and Murnighan [48], Delphi was more accurate than unstructured, direct interaction, but less accurate than a staticized group. However, the difference between Delphi and the staticized group was small. A confounding factor in this study was the big difference in the first-round estimates between the groups. They were most accurate for the Delphi and subsequently became less accurate over rounds. The unstructured, direct interacting group was least accurate on the first round and remained so. The accuracy of the staticized group was between those of the other two groups, but subsequently improved over rounds.

Three studies did not find a difference between Delphi and one or more other methods. These studies can also be criticized. In the study by Brockhoff [59], not less than seven hypotheses were tested at the same time, of which the difference between Delphi and unstructured, direct interaction was only one. Not surprisingly, a confounded result was found. Delphi was slightly more accurate with respect to forecasts and slightly less accurate with respect to fact-finding questions. As the author himself remarks, it cannot be excluded that still other factors—for instance, group size—influenced these results.

In the study by Rohrbaugh [12], both Delphi estimates and those produced in a structured, direct interaction method became more accurate over rounds, but only the last increase was significant. Rohrbaugh found no difference in last-round accuracy between both groups; this could be the result of the lower accuracy of first-round estimates in the structured, direct interacting group. A notable feature of this experiment was that participants were not instructed to give their most accurate estimates, but to "reduce the existing differences within their group." The influence of these instructions on the end result cannot be ascertained, but setting a goal other than optimal accuracy can hardly be expected to optimize accuracy.

In the study by Fischer [57], all four methods mentioned in Table 2 were compared. Amazingly, no difference between any pair of methods was found. Seaver [55] partly holds Fischer's method of analysis responsible for this. The proper scoring rules Fischer used are, according to Seaver, relatively insensitive to differences between estimates. The Delphi variant used could be called atypical. There were eight groups of only three persons each. Only two rounds were run. The minimal requirements for sharing information in a Delphi do not seem to be met in such small groups with very little opportunity for feedback.

Scrutiny of the separate studies does not lead to more easily interpretable conclusions or clues for the reasons why no comparison between methods gives an equivocal result.

The only justified conclusion seems to be that factors other than the specific method used (capability of the group leader, motivation of the participants, quality of the instructions, etc.) to a large extent determine the accuracy of an application of a judgment method. In accordance with this, one of the most consistent findings is that the method which was the primary focus of an article, and which can be expected to be preferred by the author(s), was almost always found to rank highest in accuracy [14, 20, 33, 41, 43, 49, 52, 53]. The only two exceptions are the studies by Brockhoff [59] and Rohrbaugh [12]. Brockhoff was unable to show Delphi's greater accuracy as compared to unstructured, direct interaction, and Rohrbaugh could not show his self-developed structured, direct interaction method (social judgment analysis) to be more accurate than Delphi. Both authors try to find methodological explanations for this and also mention additional advantages of their preferred method. The predominant positive results with the method of preference suggests that the unidentified factors in judgment methods responsible for high accuracy are best taken care of when the investigator has confidence in the method being used. For this reason, it is very difficult to evaluate the greater accuracy—as compared to Delphi—of a few highly idiosyncratic methods, tested only once by their developer and main proponent [10, 49, 52, 60, 61].

Contribution of the Delphi Characteristics to Its Accuracy

ANONYMITY

Anonymity in a Delphi is meant to exclude group interaction processes that decrease the accuracy of group judgment, while preserving positive influences [56, 62]. Anonymity has been criticized [63] for its intrinsic negative effects (lack of a feeling of responsibility for the end result) and because it would obviate some intrinsic positive effects of unstructured, direct group interaction (flexibility and the richness of nonverbal communication). A conclusion regarding the (dis)advantages of anonymity is difficult to draw. The only data giving a rough indication concern the satisfaction of participants with the Delphi method. This satisfaction varies strongly [48, 53, 58, 61, 64]. A possible problem with anonymity is low compliance. In one study [22], partial anonymity led to a higher response.

USE AND SELECTION OF EXPERTS

It seems obvious to use experts in situations of high uncertainty. According to many authors [45, 65–68], however, the lack of directly relevant information in uncertain situations determines judgments more than the information that is available, and consequently experts are not more accurate than nonexperts. One author [28] even contends experts perform worse than nonexperts, because the former are more strongly influenced by the desirability of an answer. Within the context of Delphi, expertise has more specifically been investigated by means of the relation between accuracy and self-ratings. Negative [42, 48, 65, 67–70] as well as positive [35, 47, 50, 71, 72] correlations have been found. Dalkey [72] tried to reconcile these contradictory results by posing that no substantial correlation exists, either positive or negative, for individuals, but that a positive correlation holds for groups from approximately seven persons. Dalkey even argues that selection of a subgroup with high self-ratings leads to the same increase in accuracy as he finds with the Delphi method. He proposes a combination of both: subgroups for questions having enough participants with high self-ratings and a Delphi for the remaining questions. A disadvantage of selecting a subgroup is that group size is reduced, with the possibility of a decrease in accuracy due to a larger random error. Research with other judgment methods has shown that selection of a subgroup increases accuracy only if the

group shows large variability in accuracy, if the probability of making a wrong judgment is great, and if the most accurate judges can be selected with high certainty [73–75].

ITERATION

The original purpose of iteration in Delphi is to have only the least-informed participants change their minds over rounds [56, 62]. Critics [45] assert iteration only leads to boredom. In a great many studies, a slight improvement in accuracy over rounds is found (for review, see Armstrong [7], Dietz [69], and Erffmeyer et al. [76]), although there are also studies finding no improvement [48, 52, 60]. In nearly all studies finding improvement, most or all improvement takes place between the first and second estimation rounds [7, 10, 20, 28, 69, 77]. In a few studies accuracy further increased after the second estimation round [60, 76]. Iteration cannot be considered independent from feedback in a Delphi. Only a few studies investigated the effects of iteration and feedback on accuracy independently from each other. In three studies [42, 48, 49], a slight increase in accuracy was caused completely by iteration, and in only one [47] by feedback.

FEEDBACK

The idea behind providing feedback is to share the total information available to a group of individual experts. Those experts who find the composite judgment of the group or the arguments of deviating experts more compelling than their own should subsequently modify their judgment [56, 62]. In this way, group pressure to conformity is prevented and any change in judgment is caused by new information only. Feedback in a Delphi can consist of the statistical summary of the group response as well as arguments of deviating judges. There is limited evidence that arguments of deviating judges leads to an increase in accuracy. Three studies [61, 71, 78] report a slight increase in accuracy due to arguments over the increase caused by statistical feedback. In one study [20] feedback of arguments decreased accuracy. In nearly all studies, the largest increase in accuracy in a Delphi is found between the first and second estimation rounds. Only the statistical group response has then been fed back.

A host of support can be found for the assertion that statistical feedback induces conformity. Numerous studies [12, 20, 28, 38, 49, 64, 79–83] show that changes over rounds are in the direction of the group response that has been fed back. For most of these studies, Ford's [49] remark holds: "Delphi methods induce change toward the median, but rarely cause the median to change." Change toward the fed-back value also occurs when this value is false [33, 80, 83, 84]. In two studies [20, 35] a "pull to the mean" has been compared directly to a "pull to the true." By far the greatest increase in accuracy was caused by the pull to the mean. In these two studies and in a later one [77], a relation was found between the distance of an individual's judgment to the group response and the percentage of judges subsequently changing their estimate (see Figure 1). This relation suggests an explanation for the slight increase in accuracy over rounds found in many Delphis. First-round estimates have been found to be distributed lognormally (see Figure 2). In such a negatively skewed distribution, the modus lies before the mean and median. Consequently, the distance to the median is on average smaller for values to the left than for values to the right of the median. Two processes may cause a subsequent change toward the median:

1. The relation between distance to the group response and subsequent change is asymptotic. Because values to the left are on average closer to the median than those to the right, newly estimated values will come even closer to the left of the old median than those to the right.

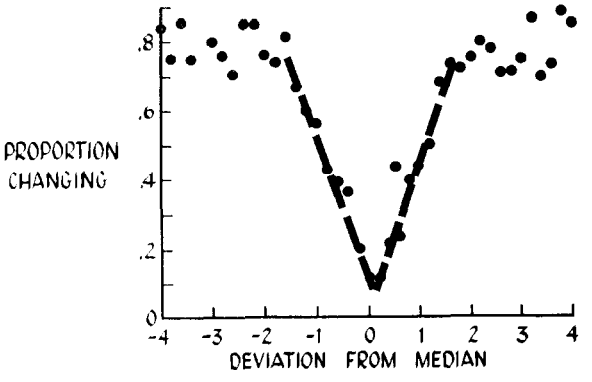

Fig. 1. Proportion of participants changing opinion versus deviation from the median. From Dalkey [20].

2. The relation between distance to the group response and subsequent change is not perfect. Also, values on and close to the median will be changed. Because values to the left of the median are on the average closer and, upon new estimation, come even closer to the median than the values to the right, there will be more crossings from left to right than from right to left.

The net result is a small increase in the median. If the initial group response underestimates the true value, the result is a slight increase in accuracy. If the initial group response overestimates the true value, the result is a slight decrease in accuracy. Support for this has been found in several studies [17, 78, 85].

A definite conclusion regarding feedback in a Delphi is now possible. The small increase in accuracy over rounds found in many Delphis can be ascribed partly to the mere iteration of judgment (see above) and partly to an artifactual by-product of the pressure to conformity caused by the statistical feedback. In any case, changes in estimates caused by feedback, whether or not associated with an increase in accuracy, are not induced by dissemination of information to all participants, but are the result of group pressure to conformity. This is supported by the finding that feedback of arguments, which can especially be though to disseminate information, has only a negligible effect on the accuracy of estimates.

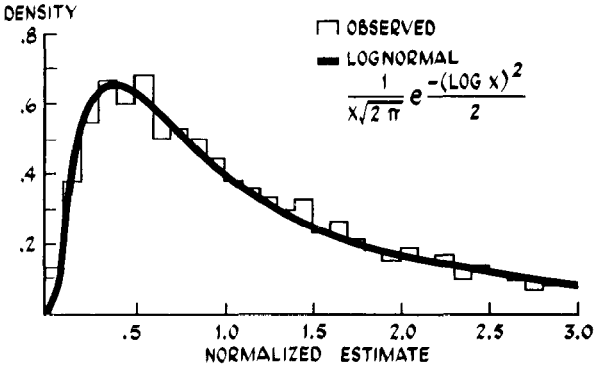

Fig. 2. Distribution of first-round estimates. From Dalkey [20].

TABLE 4
Reliability of Delphi in Studies Conducting Two or More Delphis with the Same Questions

Study	Reliability
Helmer [86]	0.87[a]
Bender et al. [88]	0.32 (−0.06–0.73)[a]
Ament [87]	0.57[a]
Dalkey [20]	0.55 (0.40–0.75)[b]
Dalkey et al. [72]	0.86 (0.59–0.98)[c]
Sahr [89]	Intuitively good
Welty [90]	Nonsignificant difference
Welty [67]	0.82[a]
Amara and Salancik [93]	0.77[b]
Grabbe and Pyke [94]	Intuitively good
Martino [see 45]	Intuitively good
Huckfeldt and Judd [95]	Intuitively good
Sackman [45]	0.70[c]
Dagenais [96]	0.52 (0.24–1.00)[d]

[a]Pearson product-moment correlation calculated by the present author.
[b]Unspecified correlation coefficient reported by the original author.
[c]Rank-order coefficient reported by the original author.
[d]Kappa reliabilities reported by the original author.

Reliability

Because of person- and situation-specific factors, as described in the section on evaluation of Delphi, it would appear difficult to standardize Delphi and to evaluate its reliability. Reliability has exclusively been evaluated by comparing results of two groups of participants in the same Delphi. Such a procedure is acceptable if it is assumed that both groups have the same bias. With proper randomization, this assumption does not appear unreasonable. Unfortunately, the studies evaluating the reliability of Delphi suffer more serious shortcomings. A drawback of many studies is that comparisons were made between groups whose forecasts were obtained in different years. Experiments were from one [86] to eight [87] years apart. Clearly, forecasting the date of a future event in 1964 is not the same as forecasting that date in 1972, when circumstances will have changed considerably. As with accuracy, many studies used no statistical criterion. Because reliability can only be inferred from absolute values and not from comparisons to other methods, as with accuracy, there is a serious shortcoming here. Therefore, if raw scores were reported, Pearson product–moment correlation coefficients were calculated. An additional shortcoming of many studies is that no variances were indicated, making it unclear whether the intuitively judged high reliability was caused by genuine similarity or by overlap of unduly wide confidence intervals. An overview of studies assessing the reliability of Delphi is given in Table 4. They all report a relatively high reliability.

Helmer [86] puts median results of three experiments next to each other. Variances were not considered. One experiment was conducted in 1963 with a traditional Delphi method [19]. Another was a pretest conducted in 1966 with 23 Rand Corporation employees. The last experiment was done at a conference using 100 conference delegates (not physically separated from one other), of which 23 were selected to finish the exercise. The article gives no indication as to how this selection was performed. The 1966 pretest and 1967 conference Delphi could be compared on seven forecasts. One event was forecasted to occur in the same year by both groups. The other six forecasts differed by 2–9 years. The correlation coefficient calculated from the raw data presented by the author was 0.87. Conference delegates remarked that serious biases existed in the wording of

the questions, and they did not expect their forecasts to be particularly "reliable." The 1963 study contained four questions related—but not identical—to those in the other two studies. Differences ranged from one to 21 years. Because of the lack of similarity and the small number of questions, computation of coefficients is pointless.

In the study by Bender et al. [88], three experiments were compared, also only with respect to median values. One was the above-mentioned study by Gordon and Helmer [19]. The other two were a pretest and the experiment proper, both conducted by the authors. The three studies could be compared on 11 forecasts. Correlation coefficients for the three pairwise comparisons, calculated from the raw data and excluding questions with unprecise answers (never, $>x$), were -0.06, 0.30, and 0.73.

In the study by Ament [87], two experiments, one the study conducted by Gordon and Helmer [19] and the other by the author in 1969, were compared on 31 forecasts. Medians as well as interquartile distances were given. The phrasing of the questions was not the same in both studies. For 13 questions, a possible bias in phrasing was indicated by the author. Limiting analysis to the 18 questions for which a possible bias was not indicated, interquartile distances overlapped for 15. Although this number seems very large, interquartile distances were also large. Excluding indefinite values, means of the interquartile distances were 22.2 years for the 1963 experiment and 13.8 years for the 1969 experiment. Corresponding means of years to forecasted occurrence were 23.7 and 20, respectively. The correlation coefficient calculated from 15 median values for which precise values in both studies were given was 0.57.

Dalkey [20] reports correlation coefficients on the basis of the median responses to 20 almanac-type questions for seven pairs of groups of different sizes. The coefficients monotonically increased with group size, being almost 0.4 for a group of two respondents to 0.75 for a group of 11 respondents.

Dalkey et al. [72] continued the above-mentioned study with an experiment involving eight pairs of groups of 15–20 respondents. Coefficients were calculated from first-round estimates. First-round scores in a Delphi are in fact staticized group responses. Dalkey et al.'s experiment therefore does not refer to Delphi at all. Another criticism of this experiment is that estimates were not used directly. Because correct answers were available, group errors were calculated by taking the natural logarithm of the median estimate of the group divided by the true answer. Dalkey's method of analysis suggests that it is instructive to look at the relation between accuracy and reliability. Textbooks often remark that reliability is a necessary condition for accuracy. But the reverse relation is also interesting. High accuracy is a sufficient condition for high reliability. Very easy questions, having high accuracy, automatically lead to high reliability. This high reliability could be called spurious. Although Dalkey did not check for this possibility, the high coefficients found in his study could merely indicate that he asked simple questions.

Sahr [89] compared three Delphi studies, but not directly with respect to reliability. Sackman [45] criticizes this study for its presentation of "some fifty pages filled with descriptive quantitative data," without reporting "a single statistic indicating variances, standard errors of measurement or product–moment reliability coefficients."

The two studies by Welty [67, 90] report intuitively high reliability. However, these studies were meant to discredit Delphi. Of the two groups compared, one consisted of experts and the other of laypeople. By showing that both groups performed equally well, Welty intended to prove that Delphi's strong reliance on experts is misplaced. The two experiments Welty performed were partial replications with lay participants of a study by Rescher [91, 92] with experts. Participants had to give estimates on a five-point scale about the importance of a number of American cultural values (14 in Welty's first study

and 17 in the second) in the year 2000. The topic makes one wonder who is to be counted as an expert and who is not. For the first study [90], Welty only reports the results of an overall sign test. This test showed both groups were not significantly different ($p = .18$). In the second study [67], an overall value could not be computed. Of the 16 questions that could be compared, only two were significantly different by means of an F-test. However, error variances were quite high. The mean of the 16 standard deviations was 0.84 for the experts and 1.13 for the layjudges. Corresponding 99% confidence intervals that go along with a one-tailed test at the 1% level (used by Welty) are 1.96 and 2.63, respectively. This leaves very little room for obtaining a difference between two groups making judgments on a five-point scale, where, in fact, 31 of 32 group scores fell between points two and four. The correlation coefficient based on 15 of the 16 pairs of means, for which values in both groups were given, was 0.82.

Amara and Salancik [93] also used an ordinal five-point scale on which two groups of participants had to rate the likelihood of 89 social developments in the 1980s. A correlation coefficient of 0.77 was found. The same criticisms as those leveled against Welty's study are valid. Most participants gave answers in the middle range of the scale, leaving very little room for a difference of more than one scale point.

In the study by Grabbe and Pyke [94], six experiments, conducted from 1964 through 1972, were compared on median forecasts. Of the six studies 15 pairwise comparisons can be extracted. The number of more-or-less comparable questions between studies ranged from zero to eight, with a mean of 2.6. Only eight comparisons involved identical questions: three concerning three questions, two concerning two questions, and three concerning one question. The small number of identical questions makes it impossible to calculate correlation coefficients for any pair of studies, therefore, nothing can be said quantitatively about the claim of the authors that "there appears to be reasonable good agreement among the dates."

An attempt to demonstrate the reliability of Delphi by Martino is mentioned and at the same time criticized by Sackman [45]. In a number of independent studies Martino noticed several similar questions that resulted in similar forecasts. No statistical indexes were reported.

Huckfeldt and Judd [95] compared the responses of five respondents answering two questions twice, separated by two weeks. Respondents had to indicate on a seven-point scale the likelihood of the occurrence of an event. The authors indicate only the percentage of responses differing one, two, three, or more scale points on the two occasions. It can be read in their data that 23% of the changes were more than one scale point and only 10% more than two points different. No rank correlation is reported and, because no raw data are given, it could not be calculated.

Along the lines followed by Welty [67, 90], also with the aim of disproving Delphi's strong reliance on experts, Sackman [45] replicated the much-referred-to Gordon and Helmer study [19] several times with students. Spearman rank coefficients were calculated for the order of years in which events were forecasted. The average rank correlation between the several replications was 0.77.

Dagenais [96] calculated kappa reliabilities for 16 pairs of groups of students asked to identify successful vocational education programs from a list of programs. Kappa values varied greatly. As with Dalkey's data, a relation with accuracy was not investigated, but could be expected. The highest kappa value Dagenais reports is 1, a value not to be expected with even the most sophisticated judgment method if something more complicated than a simple arithmetic question is asked.

The correlations reported in or derived from the above-mentioned studies range from

very low to very high. In some cases, the intuitively found high reliability could not be substantiated quantitatively. Next to the great variation in correlations and the unability to quantify some results, the discussed data show many limitations and shortcomings (relation to accuracy, bias in phrasing of questions, different years of obtaining forecast, large or unindicated error variance). A methodology to evaluate the reliability of Delphi more thoroughly has been suggested by Hill and Fowles [65]. They suggest reliability should be ascertained in several ways, including test–retest evaluation and investigation of the effects of procedural variations on the results. Their proposal has not been adopted. A definite conclusion regarding reliability of the Delphi method must therefore be postponed. The present data, however, do suggest the reliability of *the* Delphi method can hardly be expected to exist. Because of person- and situation-specific biases, a new measuring instrument is created with every new application of the method. These person- and situation-specific biases inevitably arise as every new set of questions is accompanied by a new group of experts giving judgments. Person-specific biases can only partly be removed by random selection of participants to groups. Differences between groups can also occur by selective attrition of participants during execution of a Delphi. Attrition in Delphis is very variable [45, 53, 58, 80, 95, 97]. Some data suggest attrition in a Delphi can be selective. It has, for instance, been found that dropouts are further removed from the group median on the first round than holdouts [77], implying that deviating participants drop out more often.

Consensus

In the introduction it was mentioned that in addition to the general view that interacting groups give more accurate judgments than a staticized group, interacting groups also show stronger agreement. A Delphi is extremely efficient in achieving consensus [12, 20, 28, 38, 49, 64, 79, 81–83]. Several studies report stronger consensus with a Delphi than with unstructured, direct interaction [33, 57]. As for accuracy, consensus is almost always maximum after the second estimation round [16, 18, 80, 95]. Although consensus can be important, it can never be the primary goal of a Delphi. High consensus is neither a necessary nor a sufficient condition for high accuracy. In most Delphis a slight increase in accuracy over rounds is found (see the section on iteration). In contrast, consensus increases very strongly. Also, a direct comparison shows that consensus increases much more strongly than accuracy [20]. Together with the indications that group pressure to conformity is very strong in a Delphi, this makes consensus in a Delphi suspect and in no way related to genuine agreement. Consensus in a Delphi is therefore not a good criterion, not even as a secondary one. Most probably, in situations of uncertainty lack of consensus is inevitable. As Stewart and Glantz [98] remark, "The same lack of knowledge that produced the need for a study that relied on expert judgment virtually assures that a group of 'diverse experts' will disagree."

Conclusions

The data discussed in the present article leave no other possibility open than for a negative evaluation of quantitative Delphi. The main claim of Delphi—to remove the negative effects of unstructured, direct interaction—cannot be substantiated. In many Delphis a slight increase in accuracy over rounds is found. But this increase can be ascribed partly to mere repetition of judgment (possibly by giving judges opportunity to contemplate their judgments) and partly to an artifactual consequence of group pressure to conformity. A Delphi is extremely efficient in obtaining consensus, but this consensus

is not based on genuine agreement; rather, it is the result of the same strong group pressure to conformity.

In view of the manifold applications of Delphi all over the world, the negative conclusion drawn in the present article may seem surprising. But negative evaluations of Delphi have been appearing since the 1960s. In the Rand Corporation study that aroused worldwide interest for Delphi [19], Gordon and Helmer used forecasting as a weaker term for prediction to indicate the tentative nature of their and related investigations. The fame of the study seems to be based more on the quality of the participants (Isaac Asimov, Arthur C. Clarke, Carl G. Hempel, and Stephen Toulmin, among others) than on the quality of the study itself or its results [99]. Dalkey, next to Gordon and Helmer the main developer of Delphi, summed up most of the negative aspects of Delphi, including the strong group pressure to conformity induced by statistical feedback of the group response, in two articles published in 1968 [16] and 1969 [20]. In 1971, two review articles appeared [28, 31] that for the first time prudently concluded that Delphi may not be fit for quantitative application. Not in the least prudent, this conclusion is repeated by Sackman in his well-known *Delphi Critique* [45]. Although several authors [24, 32, 84, 100] tried to rebut Sackman, critical articles about Delphi since his vehement attack have not ceased to appear [11, 77]. A few review articles favorable to Delphi have also appeared. Riggs [33] reviewed three studies and described one experiment in which Delphi was compared to other judgment methods. Delphi was the most accurate method in one of the reviewed studies, the least accurate in another, and more accurate, but not statistically significant in the third. After concluding his own study, finding a nontraditional variant of Delphi to be more accurate than unstructured, direct interaction, Riggs concludes, "in spite of the devious limitations of the study, the evidence lends credibility to the statement that Delphi procedures are superior to conference methods for long-range forecasting." Another favorable review was written by Shneiderman [61]. After reviewing five studies, he concludes that "the Delphi on the whole produces somewhat more accurate final results than the traditional committee." This conclusion seems not to be based on the reviewed studies, because of the five he reviewed, Delphi was the most accurate in two, the least accurate in two, and equally accurate to another method in the fifth.

Partly in response to the strong critique leveled against Delphi in the 1970s, the quantitative applications have become dominated by more qualitative applications (see the history of Delphi above). The proliferation of quantitative applications has been publicized by many authors [10, 26, 30, 32, 101]. They describe Delphi with the metaphor "art instead of science." It can be questioned whether the "art instead of science" approach can be regarded as a watertight defense against the critique of Delphi's many drawbacks. Quantitative and qualitative statements cannot strictly be divided. Ordering priorities or signaling future developments implies making assumptions about which social, cultural, and political developments will occur and when they will occur. In any judgment that has accuracy as a goal or, at least, underlying assumption, accuracy cannot be expelled as a necessary criterion for the particular judgment method used. Only the pure decision Delphi, where the primary goal is consensus, could perhaps be excluded from the demand of the criterion of accuracy. For all other applications, qualitative and quantitative alike, Ascher's [102] remark about forecasts holds: "accuracy is an asset because the utilization of forecasts requires, at a minimum *credibility*" (emphasis by the original author). There exist no clues that the drawbacks found in quantitative applications of Delphi do not also occur in more qualitative applications. With the negative verdict on quantitative Delphi applications, it is troubling that "the transition from forecasting and estimation to value Delphis has been lightly made without much in the way of evaluation or reanalysis" [27]. The problems this creates for Delphi are nicely illustrated in a set of statements taken

from a review by Murray [26], in which the drawbacks of Delphi are recognized, but the qualitative type of Delphi is nevertheless favored: "Delphi is thus not a science but an art," which can be used when "nothing better than opinion can be achieved," while "the final justification for the technique must be on its usefulness to decision makers." Last-resort arguments like this seem, at the least, questionable to justify the use of Delphi when it can be clearly shown that Delphi is in no way superior to other (simpler, faster, and cheaper) judgment methods.

This review was written with the support of a Dutch government grant, aimed at investigating the usefulness of judgment methods in the assessment of the toxic effects of hazardous chemicals on humans.

References

1. Dawes, R. M., and Corrigan, B., Linear Models in Decision Making, *Psychological Bulletin* 81, 95–106 (1974).
2. Hogarth, R. M., Cognitive Processes and the Assessment of Subjective Probability Distributions, *Journal of the American Statistical Association* 70, 271–289 (1975).
3. Kahneman, D., Slovic, P., and Tversky, A., eds., *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, 1982.
4. Meehl, P. E., *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press, Minneapolis, 1954.
5. Meehl, P. E., Seer Over Sign: The First Good Example, *Journal of Experimental Research in Personality* 1, 27–32 (1965).
6. Nisbett, R., and Ross, L., *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
7. Armstrong, J. S., *Long Range Forecasting: From Crystal Ball to Computer*, Wiley, New York, 1985, 116 pp.
8. Lorge, I., Fox, D., Davitz, J., and Brenner, M., A Survey of Studies Contrasting the Quality of Group Performance and Individual Performance, 1920–1957, *Psychological Bulletin* 55, 337–376 (1958).
9. Murnighan, J. K., Group Decision Making: What Strategies Should You Use? *Management Review* February, 55–62 (1981).
10. Nelms, K. R., and Porter, A. L., EFTE: An Interactive Delphi Method, *Technological Forecasting and Social Change* 28, 43–61 (1985).
11. Press, S. J., Qualitative Controlled Feedback for Forming Group Judgments and Making Decisions, *Journal of the American Statistical Association* 73, 526–535 (1978).
12. Rohrbaugh, J., Improving the Quality of Group Judgment: Social Judgment Analysis and the Delphi Technique, *Organizational Behavior and Human Performance* 24, 73–92 (1979).
13. Delbecq, A. L., Ven van de, A. H., and Gustafson, D. H., *Group Techniques for Program Planning. A Guide to Nominal Group and Delphi Processes*, Scott, Foresman and Company, Glenview, IL, 1975.
14. Erffmeyer, R. C., and Lane, I. M., Quality and Acceptance of an Evaluative Task: The Effects of Four Group Decision-Making Formats, *Group and Organization Studies* 9, 509–529 (1984).
15. Quade, E. S., Cost-Effectiveness: Some Trends in Analysis, Rand Corporation, P-3529, 1967.
16. Dalkey, N., Predicting the Future, Rand Corporation, P-3948, October 1968.
17. Kaplan, A., Skogstad, A., and Cirshick, M. A., The Prediction of Social and Technological Events, Rand Corporation, P-39, April 1949.
18. Dalkey, N., and Helmer, O., An Experimental Application of the Delphi Method to the Use of Experts, *Management Sciences* 9, 458–467 (1963).
19. Gordon, T. J., and Helmer, O., Report on a Long-range Forecasting Study, Rand Corporation, P-2982, September 1964.
20. Dalkey, N., An Experimental Study of Group Opinion: The Delphi Method, *Futures* 1, 408–420 (1969).
21. Turoff, M., The Design of a Policy Delphi, *Technological Forecasting and Social Change* 2, 149–171 (1971).
22. Rauch, W., The Decision Delphi, *Technological Forecasting and Social Change* 15, 159–169 (1979).
23. Last, J. M., Towards a Dictionary of Epidemiological Terms, *International Journal of Epidemiology* 11, 188–189 (1982).
24. Goldschmidt, P. G., Scientific Inquiry or Political Critique? Remarks on *Delphi Assessment, Expert*

Opinion, Forecasting, and Group Process by H. Sackman, *Technological Forecasting and Social Change* 7, 195–213 (1975).
25. Ludlow, J., Delphi Inquiries and Knowledge Utilization, in *The Delphi Method: Techniques and Applications*, H. A. Linstone and M. Turoff, eds., Addison-Wesley, Reading, MA, 1975.
26. Murray, T. J., Delphi Methodologies: A Review and Critique, *Urban Systems* 4, 153–158 (1979).
27. Skutsch, M., and Schofer, J. L., Goals-Delphis for Urban Planning: Concepts in Their Design, *Socio-Economical Planning Sciences* 7, 305–313 (1973).
28. Weaver, W. T., The Delphi Forecasting Method, *Phi Delta Kappan* 52, 267–271 (1971).
29. Dalkey, N., Delphi, Rand Corporation, P-3704, October 1967.
30. Helmer, O., Problems in Futures Research: Delphi and Causal Cross-Impact Analysis, *Futures* 9, 17–31 (1977).
31. Pill, J., The Delphi Method: Substance, Context, a Critique and an Annotated Bibliography, *Socio-Economical Planning Sciences* 5, 57–71 (1971).
32. Coates, J. F., In Defense of Delphi: A Review of *Delphi Assessment, Expert Opinion, Forecasting and Group Process* by H. Sackman, *Technological Forecasting and Social Change* 7, 193–194 (1975).
33. Riggs, W. E., The Delphi Technique: An Experimental Evaluation, *Technological Forecasting and Social Change* 23, 89–94 (1983).
34. Clark, A., and Friedman, M. J., The Relative Importance of Treatment Outcomes, *Evaluation Review* 6, 79–93 (1982).
35. Jolson, M. A., and Rossow, G. L., The Delphi Process in Marketing Decision Making, *Journal of Marketing Research* 8, 443–448 (1971).
36. Sahal, D., and Yee, K., Delphi: An Investigation from a Bayesian Viewpoint, *Technological Forecasting and Social Change* 7, 165–178 (1975).
37. Hays, W. L., *Statistics*, 3rd ed., Holt, Rinehart and Winston, New York, 1981, 214 pp.
38. Salancik, J. R., Assimilation of Aggregated Inputs into Delphi Forecasts: A Regression Analysis, *Technological Forecasting and Social Change* 5, 243–247 (1973).
39. Martino, J. P., The Lognormality of Delphi Estimates, *Technological Forecasting* 1, 355–358 (1970).
40. Martino, J. P., The Precision of Delphi Estimates, *Technological Forecasting* 1, 293–299 (1970).
41. Campbell, R., A Methodological Study of the Utilization of Experts in Business Forecasting, PhD dissertation, UCLA, 1966.
42. Parenté, F. J., Anderson, J. K., Myers, P., and O'Brien, T., An Examination of Factors Contributing to Delphi Accuracy, *Journal of Forecasting* 3, 173–182 (1984).
43. Pfeiffer, J., *New Look at Education: Systems Analysis in Our Schools and Colleges*, Odyssey, New York, 1968, 152 pp.
44. Sack, J., A Test of the Applicability of the Delphi Method of Forecasting as an Aid to Planning in a Commercial Banking Institution. DBA dissertation, Arizona State University, 1974.
45. Sackman, H., *Delphi Critique*, Lexington Books, Lexington, MA, 1975.
46. Brown, B., Cochran, S., and Dalkey, N., The Delphi Method. II. Structure of Experiments, Rand Corporation, RM-5957-PR, June 1969.
47. Dalkey, N., Analyses from a Group Opinion Study, *Futures* 1, 541–551 (1969).
48. Boje, D. M., and Murningham, J. K., Group Confidence Pressures in Iterative Decisions, *Management Science* 10, 1187–1196 (1982).
49. Ford, D. A., Shang Inquiry as an Alternative to Delphi: Some Experimental Findings, *Technological Forecasting and Social Change* 7, 139–164 (1975).
50. Brown, B., and Helmer, O., Improving the Reliability of Estimates Obtained from a Consensus of Experts, Rand Corporation, P-2986, September 1964.
51. Farquhar, J. A., A Preliminary Inquiry into the Software Estimation Process, Rand Corporation, RM-6271-PR, August 1970.
52. Gustafson, D. H., Shukla, R. K., Delbecq, A., and Walster, G. W., A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups, *Organizational Behavior and Human Performance* 9, 280–291 (1973).
53. Miner, F. C., Jr., A Comparative Analysis of Three Diverse Group Decision Making Approaches, *Academy of Management Journal* 22, 81–93 (1979).
54. Moskowitz, H., and Bajgier, S., Validity of the DeGroot Model for Achieving Consensus in Panel and Delphi Groups, *Journal of Interdisciplinary Modeling and Simulation* 2, 67–100 (1979).
55. Seaver, D. A., How Groups Can Assess Uncertainty: Human Interaction versus Mathematical Models. Proceedings of the International Conference on Cybernetics and Society, Washington DC, 19–21 September 1977 [Seaver, D. A., Assessing Probability with Multiple Individuals: Group Interaction versus Mathematical Aggregation. University of Southern California Social Science Research Institute: Research Report 78.3 (December 1978). Available from NTIS NON-033-0057-0967-3].

56. Martino, J. P., *Technological Forecasting for Decision Making*, North-Holland, New York, 1983, 14 pp.
57. Fischer, G. W., When Oracles Fail: A Comparison of Four Procedures for Aggregating Subjective Probability Forecasts, *Organizational Behavior and Human Performance* 28, 96–110 (1981).
58. Ven van de, A. H., and Delbecq, A. L., The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Process, *Academy of Management Journal* 17, 605–621 (1974).
59. Brockhoff, K., The Performance of Forecasting Groups in Computer Dialogue and Face-to-Face Discussion, in *The Delphi Method: Techniques and Applications*, H. A. Linstone and M. Turoff, eds., Addison-Wesley, Reading, MA, 1975.
60. Brockhaus, W. L., A Quantitative Analytical Methodology for Judgmental and Policy Decisions, *Technological Forecasting and Social Change* 7, 127–137 (1975).
61. Shneiderman, M. V., Empirical Studies of Procedures for Forming Group Expert Judgments, *Automation Remote Control* 49, 547–557 (1988).
62. Helmer, O., Analysis of the Future: The Delphi Method, in *Technological Forecasting for Industry and Government*, J. R. Bright, ed., Prentice-Hall, Englewood Cliffs, NJ, 1968.
63. Milkovich, G. T., Annoni, A. J., and Mahoney, T. A., *The Use of Delphi Procedures in Manpower Forecasting*, University of Minneapolis Center for the Study of Organizational Performance and Human Effectiveness, TR-7007 (AD747651), Minnesota, 1972.
64. Scheibe, M., Skutsch, M., and Schofer, J., Experiments in Delphi Methodology, in *The Delphi Method: Techniques and Applications*, H. A. Linstone and M. Turoff, eds., Addison-Wesley, Reading, MA, 1975.
65. Hill, K. Q., and Fowles, J., The Methodological Worth of the Delphi Forecasting Technique, *Technological Forecasting and Social Change* 7, 179–192 (1975).
66. Vinokur, A., Distribution of Initial Risk Levels and Group Decisions Involving Risk, *Journal of Personality and Social Psychology* 13, 207–214 (1969).
67. Welty, G., Problems of Selecting Experts for Delphi Exercises, *Academy of Management Journal* 15, 121–124 (1972).
68. Welty, G., The Necessity, Sufficieny and Desirability of Experts as Value Forecasters, in *Developments in the Methodology of Social Science*, W. Leinfellner and E. Kohler, eds., Reidel, Boston, 1974.
69. Dietz, T., Methods for Analyzing Data from Delphi Panels: Some Evidence from a Forecasting Study, *Technological Forecasting and Social Change* 31, 79–85 (1987).
70. Wise, G., The Accuracy of Technological Forecasts, 1890–1940, *Futures* 8, 411–419 (1976).
71. Best, R. J., An Experiment in Delphi Estimation in Marketing Decision Making, *Journal of Marketing Research* 11, 447–452 (1974).
72. Dalkey, N., Brown, B., and Cochran, S., Use of Self-Ratings to Improve Group Estimates: Experimental Evaluation of Delphi Procedures, *Technological Forecasting* 1, 283–291 (1970).
73. Einhorn, H. J., Hogarth, R. M., and Klempner, E., Quality of Group Judgment, *Psychological Bulletin* 84, 158–172 (1977).
74. Huber, G. P., Methods for Quantifying Subjective Probabilities and Multi-Attribute Utilities, *Decision Sciences* 5, 430–458 (1974).
75. Stael von Holstein, C. A. S., Probabilistic Forecasting: An Experiment Related to the Stock Market, *Organizational Behavior and Human Performance* 8, 139–158 (1972).
76. Erffmeyer, R. C., Erffmeyer, E. S., and Lane, I. M., The Delphi Technique: An Empirical Evaluation of the Optimal Number of Rounds, *Group and Organization Studies* 11, 120–128 (1986).
77. Bardecki, M. J., Participants' Response to the Delphi Method: An Attitudinal Perspective, *Technological Forecasting and Social Change* 25, 281–292 (1984).
78. Hample, D. J., and Hilpert, F. P., Jr., A Symmetry Effect in Delphi Feedback. Paper presented at the International Communication Association Convention, Chicago, 1975 (copies available from the authors at the Department of Speech Communication, University of Illinois, Urbana, IL 61801).
79. Bamberger, I., and Mair, L., Die Delphi-Methode in der Praxis, *Management International Review* 16, 81–91 (1976).
80. Cyphert, F. R., and Gant, W. L., The Delphi Technique: A Tool for Collecting Opinions in Teacher Education, *The Journal of Teacher Education* 3, 417–425 (1970).
81. Cyphert, F. R., and Gant, W. L., The Delphi Technique: A Case Study, *Phi Delta Kappan* 52, 272–273 (1971).
82. Mulgrave, N. W., and Ducanis, A. J., Propensity to Change Responses in a Delphi Round as a Function of Dogmatism, in *The Delphi Method: Techniques and Applications*, H. A. Linstone and M. Turoff, eds., Addison-Wesley, Reading, MA, 1975).
83. Nelson, B. W., Statistical Manipulation of Delphi Statements: Its Success and Effects on Convergence and Stability, *Technological Forecasting and Social Change* 12, 41–60 (1978).

84. Scheele, D. S., Consumerism Comes to Delphi: Comments on *Delphi Assessment, Expert Opinion, Forecasting, and Group Process* by H. Sackman, *Technological Forecasting and Social Change* 7, 215–129 (1975).
85. Davis, J. H., Group Decision and Social Interaction: A Theory of Social Decision Schemes, *Psychological Review* 80, 97–123 (1973).
86. Helmer, O., The Delphi Method: An Illustration, in *Technological Forecasting for Industry and Government*, J. R. Bright, ed., Prentice-Hall, Englewood Cliffs, NJ, 1968.
87. Ament, R. H., Comparison of Delphi Forecasting Studies in 1964 and 1969, *Futures* 1, 35–44 (1970).
88. Bender, D. A., Strack, A. E., Ebright, G. W., and von Haunalter, G., Delphic Study Examines Developments in Medicine, *Futures* 1, 289–303 (1969).
89. Sahr, R. C., A Collation of Similar Delphi Forecasts, Institute for the Future, WP-5, April 1970.
90. Welty, G. A., A Critique of Some Long-Range Forecasting Developments, *Bulletin of the International Statistical Institute* 54, 403–408 (1971).
91. Rescher, N., A Study of Value Change, *Journal of Value Inquiry* 1, 21–22 (1967).
92. Rescher, N., A Questionnaire Study of American Values by 2000 A.D., in *Values and the Future*, K. Baier and N. Rescher, eds., Free Press, New York, 1969.
93. Amara, R. C., and Salancik, J. R., Forecasting: From Conjectural Art Toward Science, *Technological Forecasting and Social Change* 3, 415–426 (1972).
94. Grabbe, E. M., and Pyke, D. L., An Evaluation of the Forecasting Information Processing Technology and Applications, *Technological Forecasting and Social Change* 4, 143–150 (1972).
95. Huckfeldt, V. E., and Judd, R. C., Issues in Large Scale Delphi Studies, *Technological Forecasting and Social Change* 6, 75–88 (1974).
96. Dagenais, R., The Reliability and Convergence of the Delphi Technique, *The Journal of General Psychology* 98, 307–308 (1978).
97. Dodge, B. J., and Clark, R. E., Research on the Delphi Technique, *Educational Technology* April, 58–59 (1977).
98. Stewart, T. R., and Glantz, M. H., Expert Judgment and Climate Forecasting: A Methodological Critique of "Climate Change to the Year 2000," *Climatic Change* 7, 159–183 (1985).
99. Overbury, R. E., Technological Forecasting. A Criticism of the Delphi Technique, *Long Range Planning* 1, 76–77 (1969).
100. Jillson, I. A., Developing Guidelines for the Delphi Method, *Technological Forecasting and Social Change* 7, 221–222 (1975).
101. Dijk van, J. A. G. M., Delphi Questionnaire versus Individual and Group Interviews: A Comparison Case, *Technological Forecasting and Social Change* 37, 293–304 (1990).
102. Ascher, W., *Forecasting: An Appraisal for Policy-Makers and Planners*, John Hopkins University Press, Baltimore, 1978, p. 4.

Received 26 July 1990